Improving Patient Engagement at Clinic

Zixuan Liang
Advanced Machine Learning
Northeastern University
Spring 2020

Executive Summary

Clinicas del Azucar serves a sizable population of people who live with diabetes in Mexico. The focus of the clinics is to serve the low-income population in pre-diabetes and diabetes management. We had access to datasets with clinical, demographical, and logistical information. Once we received the dataset, we performed exploratory data analysis, and hypothesis testing. Afterwards, we performed a clustering analysis using historical data: historical activity records of patients associated with the clinic. We use these records to find the engagement level of the patients over time through their patient journey at CdA. We classify the activity records of the patients moving from highest engagement level to the lowest levels, ranging from 1-12.

Once we assimilate this information of all the patients, we find the mean engagement level of the patients in each month of their patient journey. We created clusters of patients based on engagement level. The output was three clusters, with “Low”, “Medium” and “High” engagement. We narrowed our population to those who have low-medium engagement levels. We performed prediction modeling to inform variable selection in our population of interest, and proposed an intervention that would address behaviors that impact our variables of interest.

The intervention we propose involves the creation of a cellular phone application whose main goals are twofold: to improve healthy eating and physical activity behaviors in the low-engagement populations at CdA as well as to improve the engagement scores of such populations over time.

To identify features of importance, we selected a threshold value which separates the most important features from the least important after undergoing hyperparameter tuning. Then, with these features in mind, we performed causal modeling to estimate the impact of our intervention in our selected behaviors, such as coffee or refreshment intake.

As part of our future recommendations, and with increased available data to assess engagement scores, we would recommend performing ‘Clustering Analysis’ once again to find the optimal number of segments in Clinicas del Azucar to define patient groups characteristics. This would enable us to understand which patients are closely related and use this segment information for further prediction and causal analysis steps. We would like to test the effectiveness of a few more interventions such as ‘psychological behavior’, and ‘social support’ to make suggestions from various angles besides the physical and nutrients aspects.

Hypothesis for Intervention

Our team has found that there are informative variables in the data provided by Clinicas del Azucar (CdA) that can inform patient engagement over time. Variables related to eating and physical activity behaviors include clinical values (such as level of cholesterol in the blood) and can inform an intervention pertaining to such matters. Altering unhealthy behaviors (as measured by our variables) has the potential to improve clinic engagement. For our problem statement, we focus on patients who belong to low and medium level engagement groups and who have unhealthy diets (ex. consumption of sugary drinks and caffeine) and low physical activity.

Hypothesis: Patients who have unhealthy diets and are low on physical activities show lower engagement levels compared to patients who eat healthy and exercise regularly.

Hypothesis Evidence

When conducting exploratory data analysis and original variable selection, we decided on capturing variables that were collected at one point in time for most, if not all, the patients. In this way, we would ensure a sizable dataset to conduct further analyses on. A few features that were selected for the hypothesis are: IMC (BMI levels of the patients), Cafe (if patients consume coffee), Refrescos (if patients consume sugary drinks), ActividadFisica (if patients exercise or not) and Colesterol (cholesterol level of the patients). We tried to pick features that do not have a lot of null values and outliers so that we did not miss out on or have to manipulate the data. Below snippets are summary statistics performed before and after the data has been treated.

```
causal_df.isnull().sum()
```

```
IdPaciente                      0
cluster_group                   0
engagement_level                0
AnioNacimiento                  1
IdConsultorio                   0
Ciudad                         18
Estado                         16
Sexo                           83
EstadoCivil                    30
Religion                     8547
Ocupacion                     343
Escolaridad                  8630
SeguridadSocial               126
QuienPreparaAlimentos         862
EndulzanteUtilizado          1027
AlteracionesDigestivas        817
Alergias                      817
Cafe                          817
CafeTipo                     4079
Refrescos                     817
RefrescosTipo                3212
Agua                          817
ActividadFisica               817
EvaluacionDietetica          1175
Complexion                   3068
FactorActividad               817
Edad                          817
IMC                           817
Alcohol                      1088
Tabaco                       1097
Seguimiento                   870
Glucosa                      4689
HBA1C                         870
Colesterol                    870
Microalbumina                 915
TrastornosConductaAlimetaria 4952
ApoyoSocialEmocional         4952
ApoyoInstrumental            4952
TipoPersonalidad             9855
dtype: int64
```

	count	unique	top	freq
Ciudad	12279	156	MONTERREY	2869
Estado	12281	14	NUEVO LEÓN	12211
Sexo	12214	2	M	6892
EstadoCivil	12267	5	MATRIMONIO	8876
Religion	3750	101	CATOLICA	1929
Ocupacion	11954	71	EMPLEADO	4403
Escolaridad	3667	8	PREPARATORIA	919
SeguridadSocial	12171	34	IMSS	7594
QuienPreparaAlimentos	11435	4155	ELLA	1532
EndulzanteUtilizado	11270	1210	SPLENDA	2338
CafeTipo	8218	2	Cafeinado	4555
RefrescosTipo	9085	2	Light	5110
EvaluacionDietetica	11122	7859	DIETA POCO VARIADA, RICA EN CHO'S, DEFICIENTE ...	226
Complexion	9229	4	Grande	5223
Alcohol	11209	3	AUSENTE	6104
Tabaco	11200	3	AUSENTE	7960
Seguimiento	11427	1	INICIO	11427
TipoPersonalidad	2442	3	C	1017

	count	mean	std	min	25%	50%	
IdPaciente	12297.0	8561.328942	3896.148628	1.0	5638.000000	8681.0000	1183
cluster_group	12297.0	1.149549	0.502432	0.0	1.000000	1.0000	
engagement_level	12297.0	4.126164	0.503414	3.0	3.857143	4.0900	
AnioNacimiento	12296.0	1962.112557	12.446101	1900.0	1954.000000	1962.0000	197
IdConsultorio	12297.0	2.797349	1.782519	1.0	1.000000	2.0000	
AlteracionesDigestivas	11480.0	0.510801	0.499905	0.0	0.000000	1.0000	
Alergias	11480.0	0.145906	0.353027	0.0	0.000000	0.0000	
Cafe	11480.0	0.717770	0.450104	0.0	0.000000	1.0000	
Refrescos	11480.0	0.792857	0.405276	0.0	1.000000	1.0000	
Agua	11480.0	0.949216	0.219566	0.0	1.000000	1.0000	
ActividadFisica	11480.0	0.265767	0.441760	0.0	0.000000	0.0000	
FactorActividad	11480.0	0.936725	0.411893	0.0	1.100000	1.1000	
Edad	11480.0	52.573955	12.422455	0.0	45.000000	53.0000	6
IMC	11480.0	29.633685	7.180816	0.0	26.027000	29.2975	3
Glucosa	7608.0	165.138013	117.959821	0.0	102.000000	150.0000	23
HBA1C	11427.0	9.271764	3.043757	0.0	7.100000	9.3900	1
Colesterol	11427.0	155.643493	82.257433	0.0	136.700000	172.5000	20
Microalbumina	11382.0	92.001599	504.411600	0.0	0.000000	7.7000	2
TrastornosConductaAlimetaria	7345.0	0.000136	0.011668	0.0	0.000000	0.0000	
ApoyoSocialEmocional	7345.0	0.873247	0.332719	0.0	1.000000	1.0000	
ApoyoInstrumental	7345.0	0.460041	0.498435	0.0	0.000000	0.0000	

	IMC	HBA1C		Colesterol		Edad		Microalbumina	
	count	mean	std	mean	std	mean	std	mean	std
IMC									
0.0	1860	10.044855	4.026159	155.927957	83.349334	53.497312	13.630587	138.606613	729.793682
1.0	8219	9.116622	2.801728	158.284524	80.821932	52.315854	12.005515	84.166650	445.413783

	Refrescos	IMC		HBA1C		Colesterol		Edad	
	count	mean	std	mean	std	mean	std	mean	std
Refrescos									
0.0	2096	0.770992	0.420294	8.994561	3.711039	157.524857	77.907171	54.751431	12.74
1.0	7983	0.827133	0.378156	9.364944	2.894411	157.934912	82.166128	51.951647	12.15

	Colesterol	IMC		HBA1C		Edad		Microalbumina	
	count	mean	std	mean	std	mean	std	mean	std
Colesterol									
0.0	7262	30.071257	6.602987	9.121833	2.850509	52.394795	12.617710	66.657216	388.
1.0	2817	29.475077	6.614472	9.716081	3.587112	52.892439	11.547278	165.250018	731.

	Cafe	IMC		HBA1C		Colesterol		Edad		Mic
	count	mean	std	mean	std	mean	std	mean	std	mei
Cafe										
0.0	2907	0.813553	0.389534	9.288486	3.526880	156.800482	84.858055	49.440660	12.678058	102
1.0	7172	0.816230	0.387324	9.287691	2.887816	158.274888	79.807951	53.787646	11.960328	90

	ActividadFisica	cluster_group		IMC		HBA1C		Colestero
	count	mean	std	mean	std	mean	std	mean
ActividadFisica								
0.0	7452	0.757246	0.428776	30.125459	6.830709	9.391758	3.192302	0.283816
1.0	2627	0.733917	0.441992	29.278202	5.901353	8.993365	2.739654	0.267225

Cluster Analysis

Once we assimilate this information of all the patients, we find the mean engagement level of the patients in each month of their patient journey. The classes based on which we calculate the mean engagement level of the patients are stated below:

1. Cancels Membership (we don't use this class as we don't have information about this in the datasets)
2. No Engagement (absence of any records for the patient during the time period)
3. Skips Appointment
4. Cancels Appointment
5. Complains
6. Reschedules Appointment
7. Buys from Clinic
8. Attends Single-Service Appointment
9. Attends Single-Service Appointment and Buys from Clinic
10. Attends All-Service Appointment
11. Attends All-Service Appointment and Buys from Clinic
12. Buys or Renews Membership

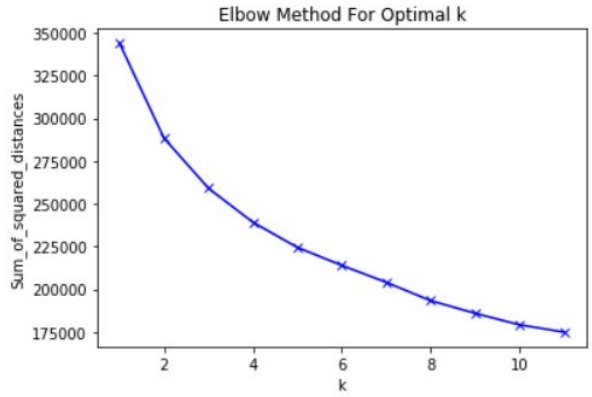
We use the class numbers (which are ordered from the lowest engagement to the highest) to calculate the mean engagement level of a patient in a particular month. When a patient does not perform any activity in a particular month, we fill in the value of the month with a 2. Below is a snippet of the dataframe that we created using this data.

	0	1	2	3	4	5	6	7	8	9	...	20	21	22	23	24	25	26	27	28	29
1	4.0	2.0	2.0	3.0	2.0	2.0	2.0	2.0	2.0	2.0	...	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
5	4.4	2.0	4.0	4.6	4.0	4.0	4.0	2.0	2.0	2.0	...	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
6	4.0	4.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	...	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
9	3.4	2.0	2.0	2.0	4.2	2.0	2.0	3.4	3.0	2.0	...	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0
14	4.6	4.0	4.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	...	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0	2.0

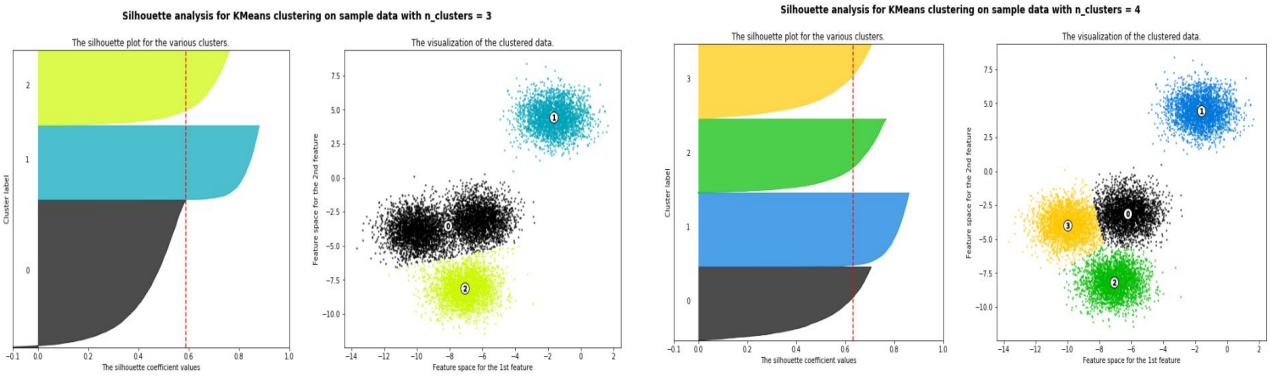
5 rows × 30 columns

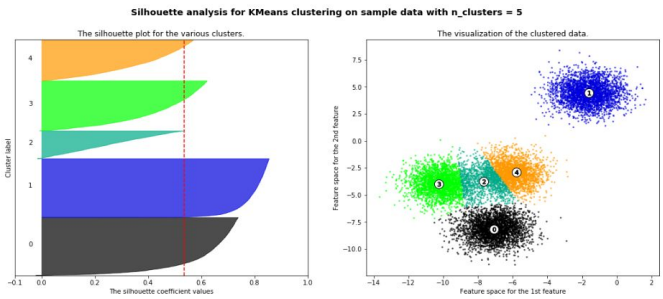
The columns here represent the months and the index values are the patient IDs. Each cell of the dataframe represents the mean engagement level of the patient and where the value is a 2.0, it means that the patient has no activity recorded for that month.

We then use this dataframe to perform the cluster analysis. We made use of k-means and GMM for the analysis. In order to determine the number of clusters, we used the elbow method and the Silhouette score. Below are snippets of the same:



From the elbow method, we cannot really determine the accurate k value. So, we move on to Silhouette analysis and find 4 is the most optimal k-value. However, for the purpose of the project we use k=3 (which does a fairly good job too).





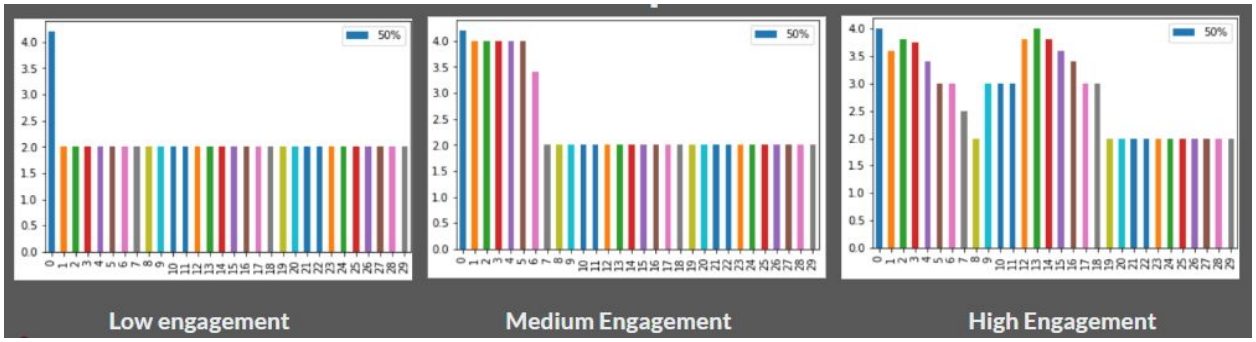
When we run a silhouette analysis visualizing 3, 4, and 5 clusters and receive the following mean silhouette scores:

For n_clusters = 3 The average silhouette_score is : 0.5875085348825803

For n_clusters = 4 The average silhouette_score is : 0.631953094161973

For n_clusters = 5 The average silhouette_score is : 0.5357177523540217

We were successfully able to distribute our patients into three cluster segments using k-means:
1. Low engagement (patients who maintained low engagement through their patient journey after the first appointment) - 8587 patients
2. Medium engagement (patients whose engagement level dropped after the 3rd appointment) - 2436 patients
3. High engagement (patients whose mean engagement level remained comparatively high to the other two clusters) - 442 patients



We also used GMM which is a soft assignment (based on the probability that a data point will belong to a cluster) approach, in contrast with k-means which follows the hard assignment approach to assign each of the data points to a cluster in order to compare the results of both the cluster analysis.
For GMM, we use k=4. The cluster segments in this type of analysis are not as structured as that in k-means. The clusters look very similar to each other irrespective of k=3 or 4. Hence, we conclude that k-means performed better.

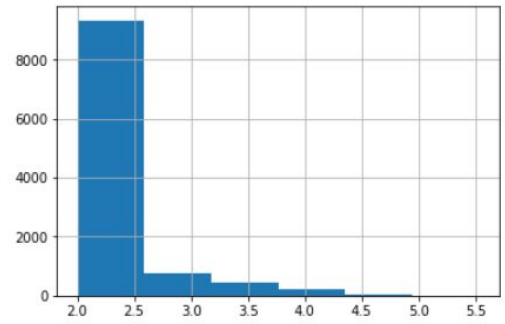
Predictive Models and Model Accuracy:

We made use of Random Forest Regression and Classification models as our predictive models to estimate the impact of an intervention around an attribute of interest on engagement outcomes. We use the first three months of the patients' activity to predict the engagement level of the 4th quarter of their patient journey. For the classification model we have 5 classes based on which we rate the engagement levels of the patients (2,3,4,5,6) and for the regression model we have a continuous range of values (2 to 5.5333).

4th_quarter_engagement_level : continous -> Regression Modeling

```
In [189]: causal_df_encoded['4th_quarter_engagement_level'].hist(bins=6)
```

```
Out[189]: <matplotlib.axes._subplots.AxesSubplot at 0x7f498d2c77b8>
```



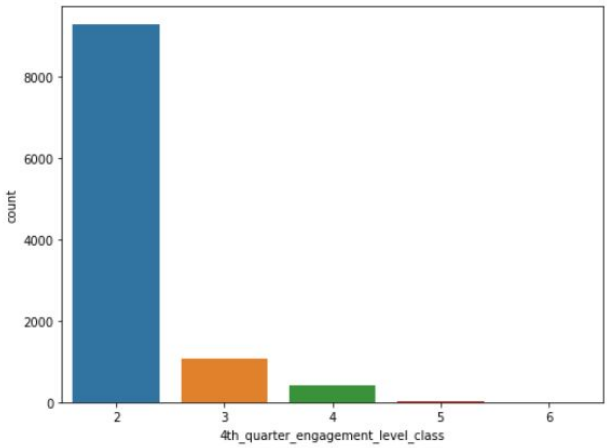
4th_quarter_engagement_level : Discrete -> Classification Modeling

```
In [190]: print(collections.Counter(causal_df_encoded['4th_quarter_engagement_level_class']))
```

```
Counter({2: 9279, 3: 1087, 4: 412, 5: 27, 6: 1})
```

```
In [191]: plt.figure(figsize=(8,6))
          sns.countplot(causal_df_encoded['4th_quarter_engagement_level_class'])
```

```
Out[191]: <matplotlib.axes._subplots.AxesSubplot at 0x7f49adda4908>
```



Sample data of the dependent and independent variables after one-hot encoding is performed:

```
In [192]: causal_df_encoded.iloc[:,1:-2].head(3)
```

```
Out[192]:
```

	cluster_group	IdConsultorio	AlteracionesDigestivas	Alergias	Cafe	Refrescos	Agua	ActividadFisica	FactorActividad	Edad	IMC	HBA1C	Colester
0	1	1	1.0	1.0	1.0	1.0	1.0	0.0	1.2	86.0	24.382	8.2	209
1	0	1	1.0	0.0	1.0	1.0	1.0	0.0	1.2	54.0	24.350	6.5	236
2	0	1	1.0	0.0	1.0	1.0	1.0	0.0	1.2	54.0	24.350	6.5	236

Microalbumina	H	M	DIVORCIO	MATRIMONIO	SOLTERÍA	UNIÓN LIBRE	VIUDEZ	Alcohol_ACTIVO	Alcohol_AUSENTE	Alcohol_INACTIVO	Tabaco_ACTIVO	Tab
0.0	1	0	0	0	1	0	0	1	0	0	0	
0.0	1	0	0	1	0	0	0	0	0	1	0	
0.0	1	0	0	1	0	0	0	0	0	1	0	

The pipeline to perform these models is as follows:

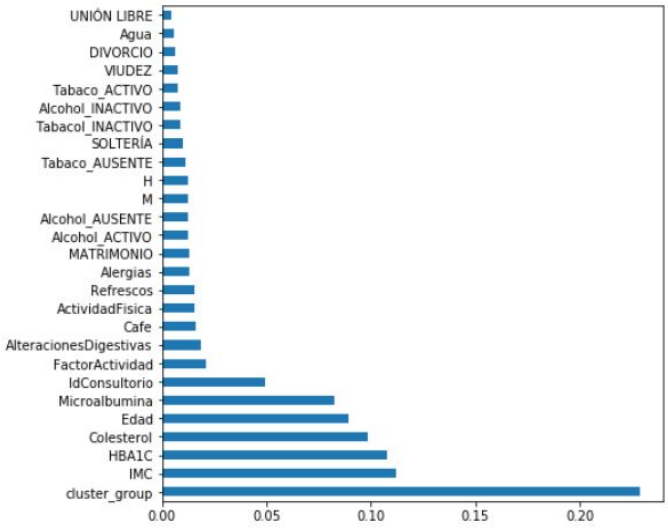
- Hyper Parameter Tuning on the Independent Variables
- Feature Importance
- Feature Selection
- Model Performance

Hyper Parameter Tuning on the Independent Variables

To understand the best combination of parameters, we use k-fold cross validation to perform Hyper Parameter Tuning on the independent variables.

Feature Importance

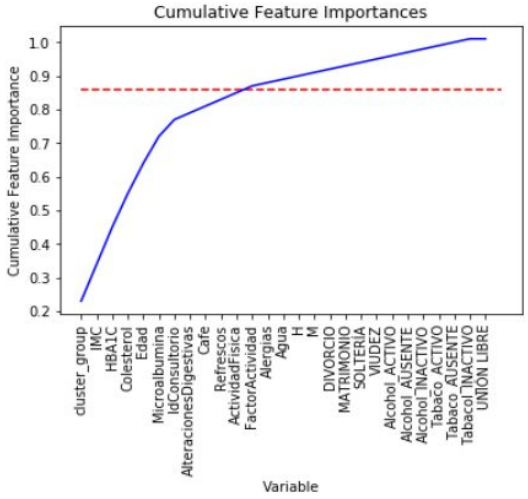
We calculate the importance of each of the features to determine which features need to be included in the model. Below is a snippet of features and their importances from the classification model.



Feature Selection

For the feature selection, we select a threshold value which separates the most important features from the least important. Then we determine the number of features and which features need to be included in the model.

For the regression model - 85% threshold; 12 features - [‘cluster_group', 'IMC', 'HBA1C', 'Colesterol', 'Edad', 'Microalbumina', 'IdConsultorio', 'Refrescos','AlteracionesDigestivas', 'FactorActividad', 'Alergias', 'Cafe’]
For the classfication model - 86% threshold; 12 features - ['cluster_group', 'IMC', 'HBA1C', 'Colesterol', 'Edad', 'Microalbumina', 'IdConsultorio', 'AlteracionesDigestivas', 'Cafe', 'Refrescos', 'ActividadFisica', 'FactorActividad']
Below is a snippet of the Cumulative Feature Importance Graph of the Classification model:



Model Performance

We then evaluate the model’s performance by calculating the accuracy.
For the regression model - We calculate the Mean Absolute Error to determine the accuracy of the regression model
For the classification model - We determine the confusion matrix from which we can observe the precision and f1 score of the classification model.
This gave model accuracy of predicting engagement level based on the features selected.
Below are snippets of the model accuracies of the Classification and Regression models respectively:

```
best_random.fit(X_train, y_train)
print("RF train feature_selected_model accuracy: %0.3f" % best_random.score(X_train, y_train))
print("RF test feature_selected_model accuracy: %0.3f" % best_random.score(X_test, y_test))
```

```
RF train feature_selected_model accuracy: 0.998
RF test feature_selected_model accuracy: 0.912
```

```
best_y_pred = best_random.predict(X_test)
print("Accuracy:", accuracy_score(y_test, best_y_pred))
print("Confusion Matrix")
print(classification_report(y_test, best_y_pred))
```

```
Accuracy: 0.9116558741905643
Confusion Matrix
              precision    recall  f1-score   support

           2       0.93      1.00      0.96      1865
           3       0.77      0.38      0.51       209
           4       0.65      0.39      0.49        82
           5       1.00      0.33      0.50         6

   micro avg       0.91      0.91      0.91      2162
   macro avg       0.84      0.53      0.61      2162
weighted avg       0.90      0.91      0.90      2162
```

```
best_reg_random = rfr_random.best_estimator_
best_reg_random.fit(X_train, y_train)
# Make predictions base_reg_random determine the error
best_reg_random = rfr_random.best_estimator_
predictions = best_reg_random.predict(X_test)
errors = abs(predictions - y_test)
# Display the performance metrics
print('Mean Absolute Error:', round(np.mean(errors), 2), 'degrees.')
mape = np.mean(100 * (errors / y_test))
best_accuracy = 100 - mape
print('Accuracy:', round(best_accuracy, 2), '%.')
```

```
Mean Absolute Error: 0.17 degrees.
Accuracy: 93.3 %.
```

The accuracy as observed from the above displayed snippets is 91.2% for the classification model and 93.3% for the regression model.

For calculating the accuracy of the Classification model, we choose a confusion matrix because it gives a generalized performance of the classes using the f1-score. By definition 'F1 score' can be interpreted as a weighted average of the 'precision' and 'recall', where an F1 score reaches its best value at 1 and worst score at 0. The relative contribution of precision and recall to the F1 score are equal. The formula for the F1 score is: F1 = 2 * (precision * recall) / (precision + recall). In this dataset, we have 5 multi-label cases, so we use the "weighted form" which calculates the average of the F1 score of each class with weighting depending on the average parameter.

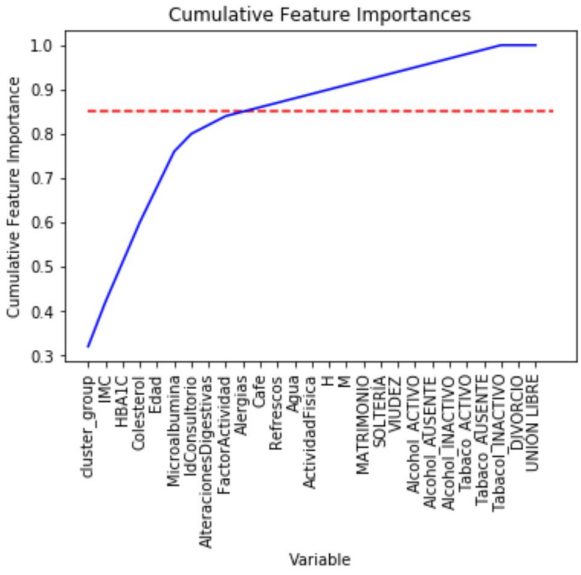
For calculating the accuracy of the Regression model, we used Mean Absolute Errors method. We use MAE for the generalized performance measure of the model. Our goal is to minimize the sum of absolute distances.

Although we performed both Regression and Classification modeling, we are more inclined towards the outcomes of the regression models because we can see the distribution in terms of continuous variables and mean engagement level difference. This is hard to observe in the classification model as the values are discrete.

Patient Population of Interest

After creating the clusters for the previous homework, we manually created subsets that would merge variables that are of interest. We decided to narrow down on our set of variables of our interest after conducting an analysis of the importance of the variables. The importance ranking ins below:

```
Feature: cluster_group          Importance: 0.32
Feature: IMC                    Importance: 0.1
Feature: HBA1C                  Importance: 0.09
Feature: Colesterol             Importance: 0.09
Feature: Edad                   Importance: 0.08
Feature: Microalbumina          Importance: 0.08
Feature: IdConsultorio          Importance: 0.04
Feature: AlteracionesDigestivas Importance: 0.02
Feature: FactorActividad        Importance: 0.02
Feature: Alergias               Importance: 0.01
Feature: Cafe                   Importance: 0.01
Feature: Refrescos              Importance: 0.01
Feature: Agua                   Importance: 0.01
Feature: ActividadFisica        Importance: 0.01
Feature: H                      Importance: 0.01
Feature: M                      Importance: 0.01
Feature: MATRIMONIO             Importance: 0.01
Feature: SOLTERÍA               Importance: 0.01
Feature: VIUDEZ                 Importance: 0.01
Feature: Alcohol_ACTIVO         Importance: 0.01
Feature: Alcohol_AUSENTE        Importance: 0.01
Feature: Alcohol_INACTIVO       Importance: 0.01
Feature: Tabaco_ACTIVO          Importance: 0.01
Feature: Tabaco_AUSENTE         Importance: 0.01
Feature: Tabacol_INACTIVO       Importance: 0.01
Feature: DIVORCIO               Importance: 0.0
Feature: UNIÓN LIBRE            Importance: 0.0
```



We narrowed our target population to people who may exhibit risky behavior regarding meals. This will show on the variables above (as these include beverage choices and laboratory values) with a threshold of 85%. This will give us insight into who may need an intervention regarding healthy eating and physical activity increase, and currently have a low level of engagement.

Creating a holistic intervention that engages the patient in a variety of ways with the clinic regarding healthy eating and exercise behaviors can improve outcomes in our variables of interest as well as engagement levels.

The variables we selected as targets for behavioral intervention (organized by level of importance) are in the table below.

Variable	**To be part of population**	**Description**
IMC	High / Low BMI We define high BMI as BMI >=25, which covers overweight and obese values	The presence of high or low BMI for each level of engagement can inform us of associations and trends on engagement. We could inform dietary interventions as well as exercise for patients with high BMI and low engagement levels.

HBA1C	High HBA1C	HBA1C informs us of the level of hemoglobin that is linked to a sugar. Patients who have higher HBA1C levels than others show a poorer control of their diabetes.
Colesterol	High Cholesterol	High levels of cholesterol in the blood may be due to inadequate diets, as well as genetic predisposition to aggregate cholesterol. As part of our study, we do not have genetic data to definitely group patients in a genetic vs dietary cholesterol aggregation group, so we will assume that high cholesterol level is due to inadequate and unhealthy diets.
Microalbumina	> 30 mg	Knowing the value of the microalbumin can give us insight into the risk for kidney failure (due an increased risk of diabetes or high blood pressure) that a patient has. ● Less than 30 mg is normal ● Thirty to 300 mg may indicate early kidney disease (microalbuminuria) ● More than 300 mg indicates more advanced kidney disease (macroalbuminuria)
IdConsultorio		We can find information depending on the clinic location--defined by the ID.
Alteraciones Digestivas	1	This is a binary--does the patient have any digestive issues?
Cafe	1	Presence of coffee in the diet (could be decaf or caffeinated)
Refrescos	1	In this case, refreshments include sugary drinks, or other beverages that are not water or coffee.
ActividadFisica	1	This is a binary variable--does the patient perform physical activities (ie. exercise, walking)
FactorActividad		This is a factor of activity based on the specific type of physical activity of the patient.

By analyzing whether patients with the presence of these variables compared to others show a lower level of engagement, we can create clinic-delivered interventions to keep patients engaged and in control of their disease progression rate.

Defining Behavior

The problem: *Patients who have unhealthy diets (incl. High sugar and caffeine beverage intake) also have low engagement levels with Clinicas del Azucar.*

The population(s): *We have over 2000 unique patients who we can cluster on the appropriate engagement levels based on our selected behavior-related predictors: BMI, HBA1C, Cholesterol, Microalbumin, Coffee, Refreshments, Physical Activity.*

The behavior and barriers: *Patients who exhibit unhealthy eating and exercise behaviors may not be doing so by free will. There may be societal barriers impeding access to healthy recipes, ingredients, and meals, as well as physical exercise. There may be also cultural influences that impact the importance that healthy eating and exercise have among the Mexican population.*

Useful Factors to consider in designing an intervention

Individual: The individual's prioritization of healthy eating and exercise. Some people prioritize healthy eating higher than others, and this intrinsic desire to eat healthily to keep the diabetes under control may be present in some subjects and not in others.

Environmental/design: The clinic seeks to serve patients in a structured manner of 1 initial long visit, and 3-month follow-ups with extra appointments for other services (such as psychology, nutrition consults). The structure of the visits and the time frames each takes may not be conducive for the strong relationship and accountability-building with patients who engage in unhealthy nutritional and exercise behaviors.

Social: Social barriers may be present in access to and maintenance of healthy eating and exercise behaviors. The ready availability and cultural importance of unhealthy foods in Mexican society may be a social barrier that is difficult to overcome. Family life and gatherings may include meals that are not in the best health-interest of the patient, but in the best interest for normalcy of the patient.

Proposed Intervention

We designed an intervention to change the behavior, for patients who have ready access to phones that can hold an application.

Disclaimer

We understand that it is likely that the low income Mexican population may not have access to cellphones that have the capabilities to hold an application. We propose the following intervention under the assumption that the patients who present at CdA and participate in the following intervention would have access to cellphones that can hold the application.

Proposed Intervention: Smartphone Application as Support Tool

Creating a human-friendly and clinic-specific application as a patient support tool to create individualized plans for healthy eating and exercise. The app would supplement in-clinic events and visits that incentivize healthy eating and exercise.

Features of the application + clinic events intervention include:

- Creation of an individualized plan that people, regardless of education level and medical background can read and follow
- Listing of tasks and milestones to be achieved
    - This would involve gamification of the meal plans: currently, patients are prescribed a meal plan after the nutritional visit, but we are uncertain of the at-home follow-up.
    - Assigning meals as tasks to be achieved could help in gamifying healthy behaviors for the patients who are struggling with low engagement. Perhaps could even help in involving family members into healthy habit formation.
- Rewards for completion the tasks/milestones
- Evaluations for BMI, HBA1C, Cholesterol, Microalbumin, Coffee, Refreshments, Physical Activity would be completed every 3 months: at touchpoints 3mo after sign-up, 6mo, 9mo and 12mo. In addition to the variables of interest, a survey asking questions about performance of the application in regard to the patient's needs and priorities is to be completed as well.
    - 12 month follow up in accordance with the already scheduled visits would enable for measurement of the variables of interest and visualization of their change over time. The visits timed at the otherwise expected checks for the typical patient would cause minimal disruption to patients schedules.
    - The survey of application performance would inform the clinic of the patient's perception towards the success of the application. This would in later analyses inform and add a layer of insight into the patient's perspective regarding the intervention and answer these questions: does perception toward the intervention go hand-in-hand with clinical figures? Does perception change over time? And finally, give feedback to the clinic on any changes the intervention may need, in order to increase success.
- The app directs the patient to engage with the clinic in a variety of ways, not only virtually. These ways include:
    - Attending check-ups regularly regardless of symptom status (hence, encouraging preventive medicine)
    - Having check-ins with the nutritionists and psychological services via telephone consult or in-person
    - Invitations and reminders for events hosted by the clinic around healthy eating and exercise, where the patient can meet people who are in the same habit-formation pursuit and create social bonds.
        - For instance: cooking workshops and recipe giveaways run by staff or volunteers.

From our cluster analysis on assignment #2, we found the following distribution of patients by cluster:

Cluster	**Patients in total**
Lowest	8577
Medium	2445

High	443

Short-term recommendations

- It is important for staff at CdA to have open conversations with their patients regarding personal priorities and goals of care. Furthermore, the staff must be willing to listen and understand that some of the patient's goals and priorities may not match those that would comprise "healthy behaviors."
- By understanding what a person values and is willing to "put in the work" for, the clinic can engage the patient with a personalized plan that involves touchpoints that have an intrinsic value to the patient.
- Patients who are part of this group of unhealthy eating and exercise and low engagement may benefit from interpersonal relationships with the staff. It is likely that variables outside of the clinic, such as social and environmental variables are barriers to their adherence.
- Explain to the patient and ask to recite back the goals of care for the next three months re: diet and exercise
- Make sure the patient understands that we are on the same team, and that CdA is more than a place deemed to "fix" the patient.
- Walk through the app with the patient slowly, to ensure the patient, regardless of education level, understands the interface and how the app fits in the overall care.

Medium-term recommendations

- Patient's use of the application must be monitored by CdA, especially during the first 3 months post personalized plan and app installation.
- In case of patients not engaging with the application, notifications could be sent (as nudges), and calls from CdA to check-in with the patient could be performed--preferably by the touchpoint at the initial intro visit.
- Keep in mind, that personal relationships with the staff are important in improving patient engagement. The staff member that introduces the patient to the intervention becomes a familiar face to the patient, as well as a support through the intervention.

Long-term recommendations

- The goal of the intervention is habit-formation. In order to create this, relationships are to be established with staff members and a support network created while the patient builds healthy eating and exercise habits.
- The app is a supplement to clinic events and visits centered around healthy eating and exercising. With this in mind, CdA would have to continue the fostering of events and relationships centered around the habit formation and incentivization.
- For patients who decrease in engagement level over time,

Notable Sources

Nezami, B. T., Lang, W., Jakicic, J. M., Davis, K. K., Polzien, K., Rickman, A. D., ... & Tate, D. F. (2016). The effect of self-efficacy on behavior and weight in a behavioral weight-loss intervention. *Health Psychology*, *35*(7), 714.

Phillips-Caesar, E. G., Winston, G., Peterson, J. C., Wansink, B., Devine, C. M., Kanna, B., ... & Charlson, M. E. (2015). Small Changes and Lasting Effects (SCALE) Trial: the formation of a weight loss behavioral intervention using EVOLVE. *Contemporary clinical trials*, *41*, 118-128.

Expected Impact on Population of Interest

We expect that patients in our population of interest, who are currently low-medium engagement, would improve in engagement score and behavior outcomes (our variables of interest). We would also expect to see the results of this intervention to come into place two quarters after the beginning of the intervention, at this point, CdA can assess the performance of the intervention and modify as needed. The life cycle for this intervention is 1 year, as we are interested in patient renewal for membership for the next year as an indicator of success.

Evaluation Metrics for the Intervention

Use of Observational Analysis for CdA Intervention Population

We would prefer analysis of **observational** data throughout the intervention. In an RCT, there is a premise that subjects will be under controlled environmental conditions throughout the trial. We understand that in the CdA population that may not be feasible, it also may be very complicated to logistically set an RCT within the clinic. There are a variety of clinics providing care to populations in Mexico, and the locations may serve people from different backgrounds and demographics, which would not enable us to easily select a truly randomized population for a trial.

By using observational data, we gain an understanding not only of the effect and causality relationships between our variables of interest, health behavior, and engagement score, but of the environment and any barriers present for healthy behavior performance.

Establishment of Causality Amongst Current Data

We use principles of matched sampling on observed data to simulate a Randomized Controlled Trial (RCT) to divide the patient population into treatment and control groups based on the behavior of each of the patients. For example, ActividadFisica - which indicates whether a patient exercises or not on a regular basis, we initialize the treatment group with patients who exercise (ActividadFisica = 1) and control group with patients who don't exercise (ActividadFisica = 0). We then calculate the 'Standardized Mean

Differences' (SMD) for feature balance detection. This is calculated by the difference in the means between the two groups divided by the pooled standard deviation. We can calculate the standardized mean differences for every feature. If our calculated smd is 1, then that means there's a 1 standard deviation difference in means. The benefit of having standard deviation in the denominator is that this number becomes insensitive to the scale of the feature. After computing this measurement for all of our features, there is a rule of thumb that is commonly used to determine whether that feature is balanced or not, (similar to the 0.05 for p-value idea).

- Smaller than 0.1: For a randomized trial, the smd between all of the covariates should typically fall into this bucket.
- Between 0.1 - 0.2: Not necessarily balanced, but small enough that people are usually not too worried about them. Sometimes, even after performing matching, there might still be a few covariates whose smd fall under this range.
- Greater than 0.2: Values that are greater than this threshold are considered seriously imbalanced.

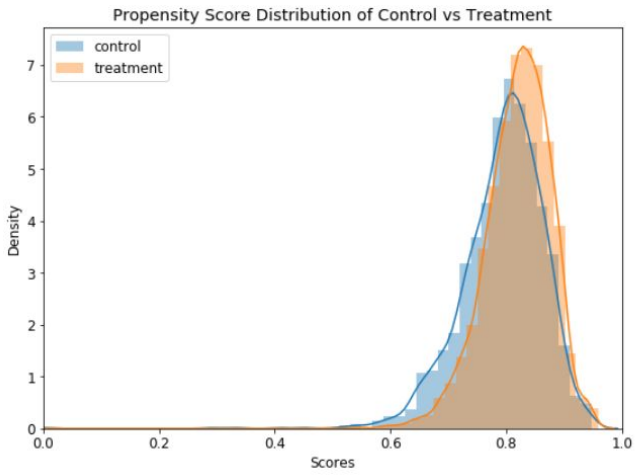
A sample for SMD calculation for all the features to perform modeling on 'ActividadFisica':

```
        smd = (pos_mean - neg_mean) / np.sqrt((pos_std ** 2 + neg_std ** 2) / 2)
        smd = round(abs(smd), round_digits)
        feature_smds.append(smd)

    return pd.DataFrame({'features': features, 'smd': feature_smds})


table_one_smd = compute_table_one_smd(table_one)
table_one_smd
```

Out[661]:

	features	smd
0	cluster_group	0.0536
1	IMC	0.1327
2	HBA1C	0.1339
3	Colesterol	0.0371
4	Edad	0.0341
5	Microalbumina	0.0629
6	IdConsultorio	0.3438
7	AlteracionesDigestivas	0.0589
8	Cafe	0.0293
9	Refrescos	0.1424
10	FactorActividad	0.0886

We used the Propensity scoring method for both Classification and Regression models to observe the differences and similarities between the distribution of the treatment and control groups. Below snippet is the Propensity score for the Classification and Regression model:



Causal effect - Comparison of potential outcomes under active treatment and control treatment, for each unit is observed by the mean engagement level difference between the two groups and by plotting the distribution of the engagement levels for the groups.

```
In [794]: control_predicted_engagement['engagement_level'].mean() - treatment_predicted_engagement['engagement_level'].mean()
Out[794]: 0.0262342045365358
```

OR

Next Steps

If we have more time and resources, first we would like to collect more historical data of the patients' engagement actions. When we calculate monthly levels of engagement, we use the scaled level of engagement from '1 to 12', which is from 'Cancel Membership' to 'Buys or Renews Membership'. However, unfortunately when we implemented the monthly engagement level matrix per patient, we found out the table was quite sparse, lacking engagement actions for the most of the months. If we can have more resources for patient engagement actions, we would like to perform 'Clustering Analysis' once again to find optimal number of segments in Clinicas del Azucar to define patient groups characteristic (demographics, physical/psychological activities, nutrients, etc) to understand which patients are closely related and use this segment information for further prediction and causal analysis steps.
Secondly, we performed 'Rubin Causal Modeling' to estimate the likely impact of the specific intervention for patient engagement such as 'intaking coffee or sugary drinks' and 'doing physical activities'. And from the findings, we would like to investigate further the effects of those specific interventions on diabetes. Also, as a further analysis step, we would like to test the effectiveness of a few more interventions such as 'psychological behavior', and 'social support' to make suggestions from various angles besides the physical and nutrients aspects.

Lessons Learned and Recommendations

Throughout the duration of the project, we learnt that:
-Validating the features that we have at hand is important in social science, especially when the outcomes are to reach populations that are vulnerable, such as low income populations.
- Prediction and causation models are complementary methods.
- We were pleased to find that Healthcare and Data Science have a strong synergy in this domain.
- We found that diabetes as a disease hasn't received enough attention and the growing number of cases, not just in Mexico, but across the world is something to be worried and active about.

According to our team, the first challenge we faced is that, since the data dictionary does not include the information of all the features present in the datasets, it is hard to grasp the exact meaning of various features. In addition, since most of the data was stored in Spanish, it was difficult to understand and process unstructured features. However, since there were a variety of variables in the database, we were able to extract more important features and perform various modeling techniques such as Clustering, Classification Modeling, Time Series and Regression Modeling without having to face a lot of difficulties.

As a team, we would like to suggest measuring the binary features currently present in the dataset, such as Social Support, Physical Activity, etc. on a scale of 1-3 (Low, Medium, High) or even 1-5, so as to make better and accurate recommendations using these features.